Kolokacje 1.1 – User's Manual Addendum

Tomasz Okniński

Contents

1 Abstract

Kolokacje is an Internet crawler and a collocation search engine. It is made up of several modules which may be executed via command-line or a graphical interface.

Usage of the program and individual modules:

```
java kolokacje.standaloneNew.SAMain [dir]
java kolokacje.standalone.SAMain [dir]
java kolokacje.crawler.Crawler dir [var1=value1 var2=value2...]
```

```
java kolokacje.index.IndexBuilder dir [var1=value1 var2=value2...]
java kolokacje.server.PrettyPrinter dir [var1=value1 var2=value2...]
java kolokacje.server.QueryServer dir [var1=value1 var2=value2...]
java kolokacje.standalone.SAManager dir [var1=value1 var2=value2...]
```

2   Definitions

In the current manual, the term *archive* indicates a user selected working directory in which the program stores downloaded texts, indices it has created, summaries and temporary files. You can place additional configuration files in the directory to adjust the behavior of individual modules.

When reference is made to a specific file in an archive, please bear in mind the fact that you can modify the configuration and rename the file in question. However, this is rather uncommon, so the manual refers to default file names, e.g. `addresses.acc`, instead of *file whose name is defined by the addressFile variable*.

3   New module StandaloneNew

3.1   Short overview

This module allows the user to work over an archive with a graphical interface. It requires access to a graphics mode provided by some kind of window manager (e.g. MS Windows or KDE, Gnome or similar under Linux).

3.2   Usage

Synopsis:

```
java kolokacje.standaloneNew.Main [dir]
```

Examples:

```
java kolokacje.standaloneNew.Main
java kolokacje.standaloneNew.Main emacs
```

`dir` – archive directory defined in 2.

3.3   Program menu

3.3.1   File

Allows the user to create and open archives and save shown results.

New archive – Shows a simple wizard that allows for creating a new archive and start crawling immediately at provided URL.

Open archive – Provides means to open required archive. Creating a new archive directory is also possible.

Save results from active tab – Saves the results shown on currently selected tab. If there is no tab in the main window, this option is not available.

Preferences – Program settings window.

Recently opened archives – A list of up to 5 recently opened archives providing quick access to them. This list may be found in `recents.ini` file. If this file doesn't contain any entries, the list is empty.

Exit – Closes program. If the index wasn't saved after the last modification, the user is provided the means to do it now.

3.3.2 Crawler

This menu controls the Crawler. If no archive is selected, items in this menu are unavailable.

Enter URL – Provides a field to enter URL which will be used as the root for the crawling process.

Start crawling – Starts crawling.

Delete old content – After starting the crawler previous content of the archive will be deleted.

Resume interrupted – Allows resuming the process of crawling from where it was stopped by the user.

View entered URLs – Creates a new tab with addresses entered by the user. This option is unavailable if no URL was entered yet.

View indexed pages – Creates a new tab with addresses indexed by the crawler. This option is unavailable if no URL was indexed yet.

3.3.3 Index

This menu allows the user to build and save the index of the archive. If no archive is opened, these functions are not available.

(Re)Build index – Builds a new archive index.

Save index – Saves built index. Not available if the index hasn't been built yet.

(Re)Build and save index – Builds and then saves the archive index.

Index configuration – Provides access to IndexBuilder settings:

- segment separators,
- word separators,
- ignorable words,
- minimal accepted word length.

3.3.4 Query

Allows the user to ask queries to the archive. Filters used by different query types are listed in table 1.

Find words – Creates a new tab showing words that match the pattern entered in the *Query* field.

Find collocations – Creates a new tab showing collocations that match the pattern entered in the *Query* field.

Find sample contexts – Creates a new tab showing contexts for words/collocations that match the pattern entered in the *Query* field.

List words – Lists all words found in the archive. This option is equal to executing the **Find words** function with an empty pattern.

3.3.5 Collocation tests

This menu is used to select which collocation tests should be used in collocation rankings.

3.3.6 Options

Provides access to different filters used by the query functions. These functions and their filters are listed in table 1.

Word order – Words sorting:

- alphabetical,
- frequency based,
- a tergo.

Include numbers – If numbers should be included in results.

Proper names – Proper names in results:

- include,
- exclude,
- only.

Collocations – Types of shown collocations:

- both left and right,
- only left,
- only right.

Minimal number of occurences – Number of required occurences which qualifies certain word/collocation for being included in query results.

Sample context size – Length of shown left and right contexts.

Table 1: Filters used for different queries

Query type	Used filters
Word list	Word order
	Numbers
	Proper names
	Min. number of occurences
Collocations ranking	Collocation test to use for sorting
	Numbers
	Proper names
	Collocations type
	Min. number of occurences
Ranking comparison	Numbers
	Proper names
	Collocations type
	Min. number of occurences
Sample contexts	Sample context size

3.3.7 Help

Shows information about the program.

3.4 Toolbars

Allow access to fields and buttons for quick querying. They are disabled if no archive is currently opened. Toolbars contain:

- Upper bar

  1. query field,
  2. "find words" button,
  3. "find collocations" button,
  4. "find sample contexts" button,

5

5. expandable menu to select which test should be used for sorting collocations query results.

- Lower bar

  1. word order selection,
  2. minimal number of occurences which qualifies a word/collocation for being included in query results,
  3. type of shown collocations.

3.5 Panel with tabs

Tabs with query results are shown inside the main program window.

3.5.1 Tab headers

Each tab header includes:

- buttons to close or detach the current tab,

- type of shown results,

- query that was used to generate shown results,

- codes of filter settings that were used for generating shown results (see table 2).

3.5.2 Popup menu

Each tab provides a popup menu, which can be brought up by clicking the right mouse button on the tab header. This menu allows for:

- showing informations about filter settings used for generating results shown on the current tab,

- redoing the query for new settings,

- switching to a tab which was the source of query for the current tab,

- saving results,

- detaching the tab from the main window,

- closing the tab.

Availability of different functions is based on current program state and type of shown results.

Table 2: Filters and their codes

Filter	Code
Word order – alphabetical – frequency based – a tergo	 A F T
Collocations sorting	 ranking used
Numbers – include – exclude	 1 0
Proper names – include – exclude – only	 n - N
Collocations type – both left and right – only left – only right	 B L R
Min. number of occurences	 number
Context size – left – right	 number number

3.5.3 Changing pages of results

Tabs support changing pages of results by means of clicking buttons located in the top-right corner of the frame:

- "skip to first page",

- "skip some pages backwards",

- "skip to previous page",

- "skip to next page",

- "skip some pages forwards",

- "skip to last page".

3.6 Additional information

3.6.1 Archive querying

Complete informations about querying archives can be found in the original version of the User's Manual in section 6.

3.6.2 Program preferences

Program preferences are described in the original version of the User's Manual in section 11.

However, some additional settings have been made available in the *config.ini* file:

Variable	Default value	Meaning
locale	pl_PL	interface language
enteredUrlsList	entered.url	file which contains URLs entered by the user
defaultNewRowsPerPage	200	default number of lines shown on single page of results in the standaloneNew version of the interface
defaultNewSkipPages	10	number of pages to skip by fast forward/backward button in the standaloneNew version of the interface
logFile	kolokacje.log	file which stores the program log
logActions	yes	if user actions should be logged
recentArchivesFile	recent.ini	file which stores 5 recently opened archives